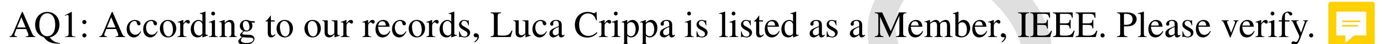
AUTHOR QUERIES

AUTHOR PLEASE ANSWER ALL QUERIES

PLEASE NOTE: We cannot accept new source files as corrections for your paper. If possible, please annotate the PDF proof we have sent you with your corrections and upload it via the Author Gateway. Alternatively, you may send us your corrections in list format. You may also upload revised graphics via the Author Gateway.

Carefully check the page proofs (and coordinate with all authors); additional changes or updates WILL NOT be accepted after the article is published online/print in its final form. Please check author names and affiliations, funding, as well as the overall article for any errors prior to sending in your author proof corrections. Your article has been peer reviewed, accepted as final, and sent in to IEEE. No text changes have been made to the main part of the article as dictated by the editorial level of service for your publication.



AQ1: According to our records, Luca Crippa is listed as a Member, IEEE. Please verify.

AQ2: Please confirm or add details for any funding or financial support for the research of this article.

AQ3: Please confirm if the location and publisher information for Reference [33] is correct as set.

IEEE Proof

Investigating 3D NAND Flash Read Disturb Reliability With Extreme Value Analysis

Cristian Zambelli, *Member, IEEE*, Luca Crippa, *Member, IEEE*, Rino Micheloni, *Senior Member, IEEE*, and Piero Olivo

***Abstract*—The storage systems relying on the 3D NAND Flash technology require an extensive modeling of their reliability in different working corners. This enables the deployment of system-level management routines that do not compromise the overall performance and reliability of the system itself. Dedicated parametric statistical models have been developed so far to capture the evolution of the memory reliability, although limiting the description to an average behavior rather than extreme cases that can disrupt the storage functionality. In this work, we validate the application of an extreme statistics tool, namely the Points-Over-Threshold method, to characterize the read disturb reliability of a 3D NAND Flash chip. Such technique proved that the die reliability characterized through extreme events analysis can be predicted using a low number of samples and generally holds good prediction features for distribution tail events.**

***Index Terms*—3D-TLC NAND flash, read disturb, reliability, points over threshold.**

I. Introduction

MODELLING the reliability of the 3D NAND Flash memory technology [1], [2] is still an important task to be performed as a support for storage system designers dedicated to Solid State Drives (SSDs) or Multi Media Card (MMC) products development. Indeed, all the firmware solutions implemented in their controllers (i.e., the computing core of SSDs and MMCs) whose goal is mitigating the inherent bit error rate (BER) exposed in different storage working conditions (e.g., endurance stress, data retention at high temperature, etc.) are well founded on memory reliability models [3]. To this extent, dedicated parametric statistical models [4]–[6] have been developed so far to capture the evolution of the memory's errors distribution through well-known statistical frameworks (defined as probability distributions) like Gaussian, Binomial, Poisson, Gamma, and so on. However, the large process-induced variability of the error characteristics in 3D NAND Flash devices [7] combined with an intrinsic difficulty in testing all the possible permutations of the memory working corners during lifetime and on a relevant statistical population, could hamper an accurate description of the distribution upper tail. It is worth to mention that on this part of the errors distribution, the storage system designers spend a significant effort to tailor the Error Correction Codes (ECCs) [8] strength and the secondary correction schemes like soft decoding [9], [10], Moving Read References [11], and even RAID [12]–[15]. Therefore, the more precise and accurate is the model the less is the probability to incur in storage performance slowdowns due to improperly calibrated error correction techniques [16].

In [17], we proposed for the first time to apply a parametric model commonly exploited in extreme value analysis (EVA) for natural sciences and econometrics to 3D NAND Flash errors distribution upper tail modeling, namely the Point Over Threshold (POT) method [18]. The work demonstrated the potential of this methodology combined with the Generalized Pareto Distribution (GPD) in the analysis of the read disturb stress after data retention. We picked that memory working condition since it is known to represent a critical use case in data center applications for Big Data analytics performed on cold data [19]. Starting from our previous work, we extended the study in twofold directions: the first related to the characterization and estimation of the read disturb also for hot data scenario (i.e., read after many updates) mimicked by an endurance stress, showing that it is not critical for the reliability as in retention conditions; the second concerning the cross-validation of the POT and the extension of its implementation with a three-parameters Weibull distribution. This will demonstrate, with a proper confidence, that POT is a powerful statistical tool to estimate the 3D NAND Flash die reliability.

II. Experimental Setup

A. Devices Under Test

The experimental activity in this work is based on the characterization of an off-the-shelf sub-100 layers 3D NAND Flash memory product implementing the Triple Level Cell (TLC) paradigm (see Fig. 1). Such technology is considered, based on its endurance and retention rating, as a mass storage medium for enterprise SSD applications. The statistical sample under investigation is composed by all the pages in every physical layer of 40 memory blocks distributed on multiple dies and chips to account for process-induced variability [7]. Since we are testing a TLC memory, we consider all the page types in the analysis (i.e., LSB-Least Significant Bit page, CSB-Center Significant Bit page, and MSB-Most Significant Bit page). Each page is sized 16 Kbytes plus the spare bytes exploited for

Manuscript received August 2, 2021; accepted August 27, 2021. *(Corresponding author: Cristian Zambelli.)*

Cristian Zambelli and Piero Olivo are with the Dipartimento di Ingegneria, Universita degli Studi di Ferrara, 44122 Ferrara, Italy (e-mail: cristian.zambelli@unife.it).

Luca Crippa and Rino Micheloni are with the Flash Signal Processing Labs, Microchip Corporation, 20871 Vimercate, Italy.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TDMR.2021.3108941.

Digital Object Identifier 10.1109/TDMR.2021.3108941

1530-4388 © 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

IEEE Proof

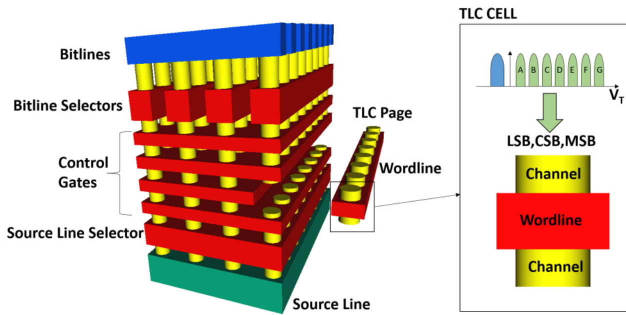


Fig. 1. TLC 3D NAND Flash architecture considered in this work [17].

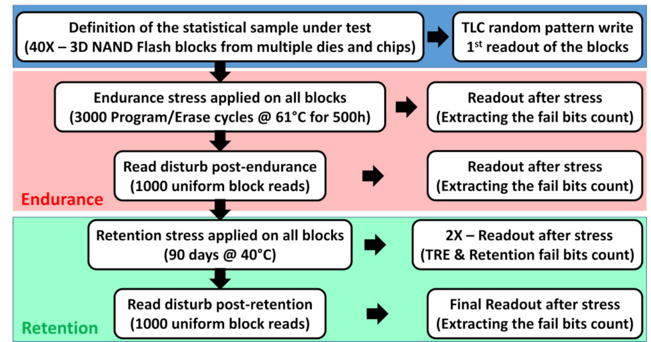


Fig. 2. Depiction of the test flow adopted in this work for the 3D NAND Flash reliability characterization.

data recovery purposes in case of data corruption. However, state-of-the-art ECCs implementations [8] work on subset of the page dimension, generally referred as a codeword (CW). In this work, the CW size is 4 Kbytes plus the spare bits, so that every page is constituted by 4 CWs. The statistical sample under test is constituted by 184320 CWs.

B. Test Flow for Reliability Assessments

The execution of the test flows required to extract the experimental data to fit with the statistical models presented in the paper is performed by the automated test equipment (ATE) presented in [20]. The system interfaces with the 3D NAND Flash chips at a 400 MT/s data rate and allows, for any applied test, to measure the number of corrupted bits (also known as fail bits count or errors number) in each CW. We remind that a bit is considered corrupted after reading from the 3D NAND Flash under test if its value changes from what has been previously written as a result of a reliability degradation process. Fig. 2 summarizes the test procedure performed. After the definition of the statistical sample under test, we write a random pattern on all the TLC pages (therefore on all the CWs) to rule out any topological dependency of the corrupted bits and perform a readout of the memory content. Then, the devices are submitted to an endurance stress up to the rated endurance of the technology (i.e., maximum number of sustainable block erase before unrecoverable errors) using a JEDEC-based cycling test [21]. The test consists in 3000 Program/Erase cycles at a 61 °C temperature for 500 hours. After the stress we perform a readout of all the CWs under test and we extract the number of corrupted bits for each CW. A read disturb is then performed post-endurance by employing a 1000 uniform block reads access pattern [22] on all tested blocks and performed another readout for fail bits count extraction. Immediately after the end of the read disturb post-endurance stress, a data retention stress is performed by placing all the devices under test in idle for 90 days at a 40 °C temperature. A double readout is performed at the end of the test to separate the Temporary Read Errors (TRE) effect typical of 3D NAND Flash architectures [23] from the retention stress results. Finally, an additional read disturb post-retention is performed and the fail bits count per CW are extracted accordingly.

III. 3D NAND Flash Read Disturb Characterization

We started the read disturb characterization on our devices by evaluating the Empirical Cumulative Distribution Function (ECDF) of the fail bits count extracted on all the CWs before and after the read disturb stress post-endurance and post-retention scenarios, as described in Fig. 2. For a given number of fail bits equal to t in a sample x, the ECDF is defined as the proportion of the values in $x \leq t$. We prefer to use the term ECDF rather than CDF since the latter one can be misleading as it is usually referenced to a theoretical probability distribution used to fit the experimental data, which is not our case here. Unfortunately, due to confidentiality reasons on the tested 3D NAND Flash samples we cannot disclose the ECDF of the fail bits count, but we have to normalize the number of corrupted bits on a CW to a defined entity. In this work, we normalized the fail bits count with respect to the ECC capacity offered by an advanced correction engine that incorporates secondary error correction schemes (e.g., read retry and soft-decoding) as well [9], [10], [24].

Fig. 3 shows that after endurance stress the read disturb generally increases the number of errors as expected from other studies in this context [25], [26]. This behavior is related to an over-programming of the memory cells that are not involved by the read operation due a moderate voltage applied to unselect them [27]. Looking at the median of the ECDFs per page type it is observed that the CWs belonging to MSB pages are those displaying the largest errors increase, even if there are some CWs in LSB pages (those in the associated ECDF tail) that are largely affected. However, the read disturb post-endurance is not particularly detrimental for the reliability since the ratio fail bits count/ECC capacity is well below one and displays a sufficient margin for safe operation without data corruption.

Fig. 4 shows the results of the same analysis replicated for the retention domain. In the errors analysis we also reported the ECDFs retrieved on the first readout of the memory blocks under test to evaluate the impact of the TRE, as discussed in the Section II of this work. We interestingly observe that the read disturb applied post-retention can recover part of the errors in a CW for all TLC page types. This has been explained in [26], by a charge redistribution mechanism occurring during

IEEE Proof

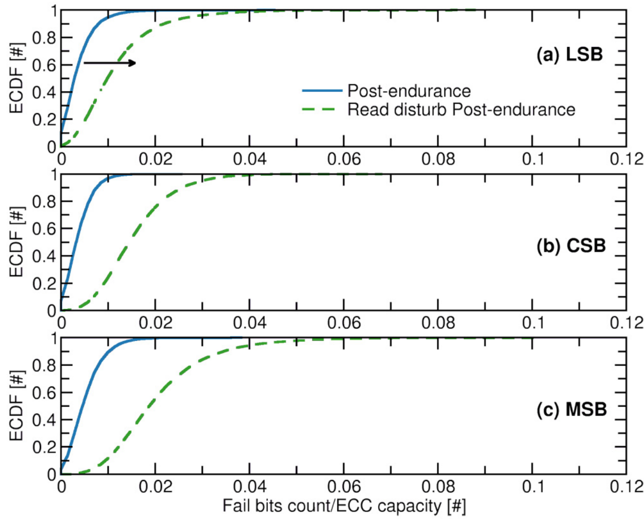


Fig. 3. ECDFs per TLC page type of the fail bits count normalized with respect to the ECC capacity in post-endurance stress and after the application of the read disturb. The ECDFs are extracted from all tested 3D NAND Flash CWs.

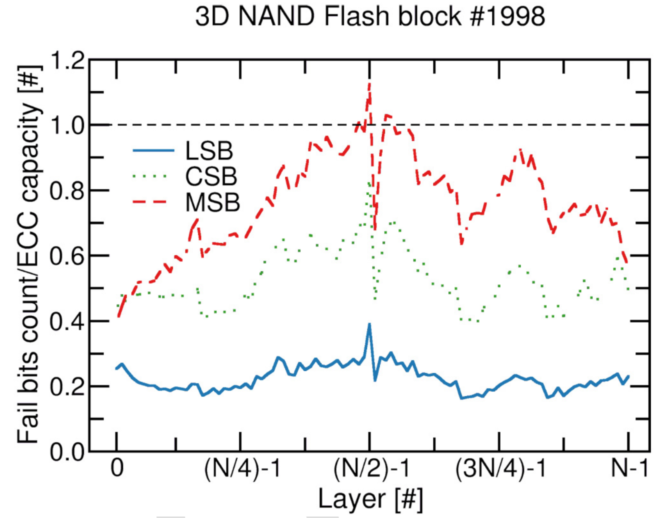


Fig. 5. Fail bits count normalized with respect to the ECC capacity characteristics per TLC page type as a function of layer position in a 3D NAND Flash block after read disturb post-retention. The ECC limit is highlighted for clarity with the black dashed line.

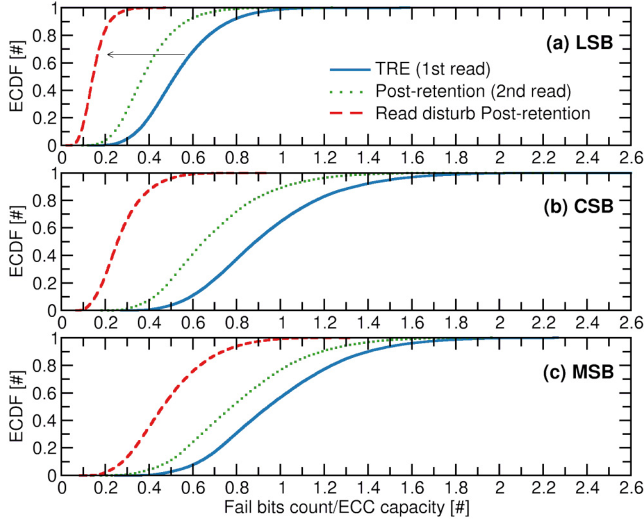


Fig. 4. ECDFs per TLC page type of the of the fail bits count normalized with respect to the ECC capacity in post-retention stress and after the application of the read disturb. The effect of the TRE is also evidenced in the figure.

the read operation. From the reliability standpoint, we note that the retention scenario is the most critical to address since the ratio fail bits count/ECC capacity can be greater than one (especially for MSB pages), thus hampering the data recovery operations.

Besides the errors' distribution characterization according to the TLC page type, we analyzed what is the contribution of the topological position (i.e., the layer position in a 3D NAND Flash block) on the fail bits count after the application of the read disturb. Fig. 5 shows the results of this investigation for the read disturb post-retention test case. We focus on this stress condition since it is the one triggering the highest number of errors during tests. We can note a large error variability among layers in a single 3D NAND Flash block after the post-retention read disturb and on top of this, there is a large error variation between LSB pages and CSB/MSB pages. This is a critical aspect that should be tackled by a statistical model developed to capture errors characteristics. Finally, to better understand the role of the read disturb on the errors count and therefore on the memory reliability we have calculated the error amplification (EA) factor as:

$$EA(i) = \frac{ERD(i)}{EPRE(i)} \quad (1)$$

where i is the layer position in the block from 0 to $N-1$ with N the number of layers exploited in the manufacturing of the 3D NAND Flash chip, *ERD* is the number of errors post-read disturb and *EPRE* the number of errors pre-read disturb. The EA is calculated both for endurance and retention test cases. Even if the endurance case is the one showing the largest EA, we must report once again that such scenario is not critical for the memory reliability since the errors amount remains always well below the ECC limits. On the contrary, even if the EA is below unity for post-retention read disturb, there are some critical conditions on which the errors count is higher than the ECC capacity and are worth to be modeled for future reliability considerations.

IV. Standard Statistical Modeling Approach

The common procedure adopted for modeling the fail bits count, and in general for all the parametric statistical frameworks applied to 3D NAND Flash data, is to fit the entire ECDF retrieved in a specific working condition. For our study case this is after a read disturb stress performed after endurance or after data retention at high temperature. The advantage of such parametric approach is to achieve a rapid estimation of the ECC capacity to cover errors and proven to be useful in many cases. The statistical modeling of CWs fail bits count distribution bases on the assumption that the corrupted bits in a CW are treated as independent events. By considering a CW length of n bytes, it is possible to calculate

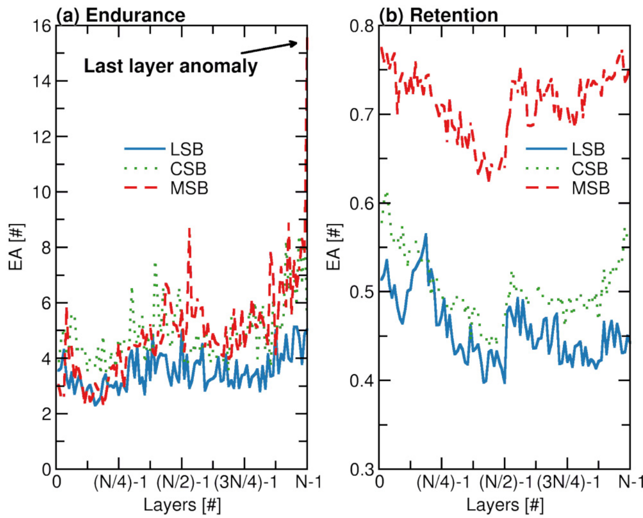


Fig. 6. EA factor for read disturb calculated as a function of the layer position in the memory block in endurance (a) and retention (b) test cases.

the probability of having k errors in the CW exhibiting a BER p using the binomial distribution probability density function as in [4]:

$$y = P_{error}(k|n, p) = \binom{n}{k} p^k (1-p)^{n-k} \tag{2}$$

However, considering that in 3D NAND Flash technology n (equal to 4 Kbytes plus spare bits in our work) is relatively larger than k and p is usually lower than 2×10^{-2}, the binomial approach starts to fail. Some alternative distributions like the beta-binomial [28], the Gamma [29], the Gamma-Poisson compound [6], or the Weibull [30] have been in consideration by the literature due to their capability in accounting the intrinsic variability of the memory technology. The easiest to calculate with software tools for numerical analysis are the Gamma and the Weibull distributions. The former is based on the following probability density function:

$$y = P_{error}(\lambda|\alpha, \beta) = \frac{\lambda^{\alpha-1}}{\Gamma(\alpha)\beta^{\alpha}} e^{-\left(\frac{\lambda}{\beta}\right)} \tag{3}$$

where λ is calculated as $n \cdot p$, α is the shape factor, and β is the scale factor of the distribution. The latter is:

$$y = P_{error}(k|\alpha, \beta) = \frac{\beta}{\alpha}\left(\frac{k}{\alpha}\right)^{\beta-1} e^{-\left(\frac{k}{\alpha}\right)^{\beta}} \tag{4}$$

with α and β being the shape and the scale factor of the distribution, respectively.

Unfortunately, due to the extreme variability characteristic of the fail bits retrieved in different locations of a 3D NAND Flash chip (see Fig. 5), both statical models (Gamma and Weibull) do not pass the χ^2 goodness-of-fit test (p-value = 0). One may still argue that even if the models do not pass the test, they are still valid for predictions of the ECDF distribution tails, whose practical applications are the selection of the ECC capacity to cover errors or the evaluation of the reliability margin. To this extent, we run a cross-validation test using a Holdout methodology where 70% of the CWs tested in the experiments are used for training the statistical models and the remainder 30% are used for testing its prediction accuracy. On a total of 1000 cross-validation splits, none of them passed the goodness-of-fit test. Analyzing the histogram count (see Fig. 7) of the fail bits count/ECC capacity and the resulting fits of the statistical models, we evidence an underestimation/overestimation of the empirical data distribution, possibly hampering the selection of the correct ECC capacity or its margin. The Gamma distribution performs better with respect to the Weibull, but still do not pass any statistical test. Same considerations can be drawn by looking at the Cumulative Distribution Function (CDF) modeled by the two statistical models. As in 3D NAND Flash reliability modeling we are mainly interested in the low probability tail of the errors distribution (i.e., extreme events), we need a framework capable to handle only that part of the empirical data.

V. The POT Method for Read Disturb EVA

A. Introducing the POT-GPD

Motivated by this, we explored a statistical framework related to EVA [18] that is commonly used in tail data analysis, namely the POT. By considering each 3D NAND Flash CW as a sequence of *i.i.d.* measurements $x_1, x_2, \ldots, x_n$, we can define as extreme events all the CWs that exceed a defined error threshold u for which we can define an exceedance as:

$$\{x_i : x_i \geq u\}. \tag{5}$$

If the exceedances are labeled as $x_{(1)}, \ldots, x_{(k)}$, it is possible to define a threshold excess as:

$$y_j = x_{(j)} - u \qquad j = 1, \ldots, k. \tag{6}$$

From the probability theory it is proven that a random variable Y_i based on the threshold excesses follows a GPD [18]. For a large enough threshold u, we can write its probability density function as:

$$f(y) = \sigma^{-1}\left(1 + \frac{\xi y}{\sigma}\right)^{-1-\xi^{-1}} \tag{7}$$

with the parameters ξ and σ being the distribution shape and scale factors, respectively.

B. Threshold Choice

The most critical operation in POT-GPD statistical modeling is the extrapolation of the best threshold to apply. The selection of an optimal threshold within a region of interest (ROI) requires a bias-variance trade-off and a knowledge of the ECC capabilities offered in the 3D NAND Flash data recovery processes. If the chosen model threshold is too low, the results are biased because of the model asymptotic assumption being invalid. In other words, a too low threshold will result in having exceedances not converging to the GPD, since the probability distribution is based on its capability of fitting extreme events [31]. On the other hand, if the threshold is too high, the variance is large due to few exceedances. In [32], it is stated that the threshold must be high enough for the exceedances over threshold to converge to the GPD, while the sample size should be large enough to ensure that there are enough data

IEEE Proof

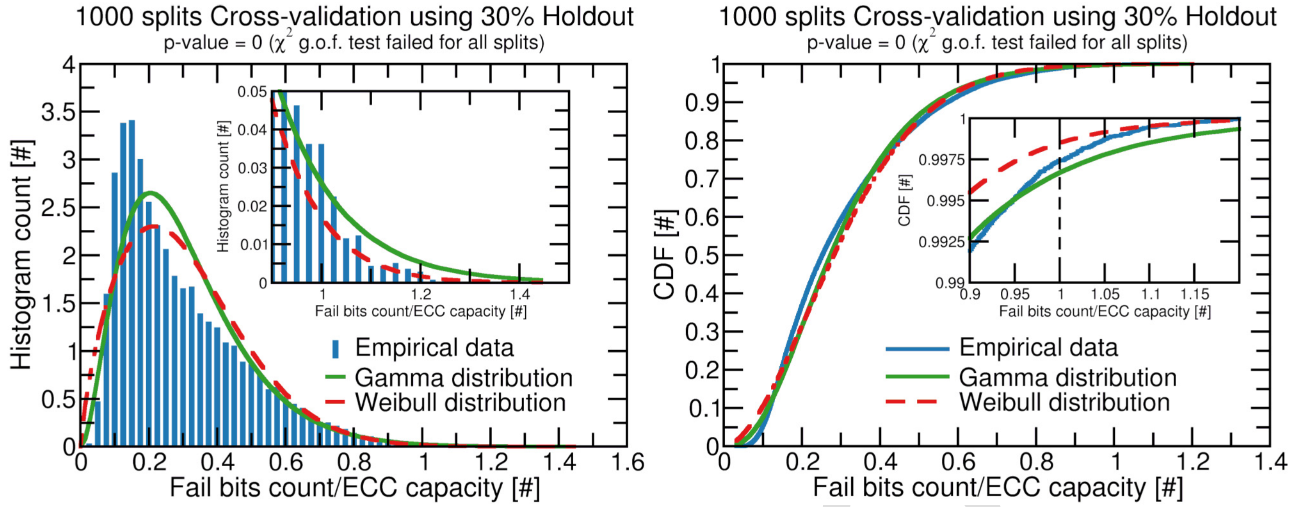


Fig. 7. Gamma and Weibull distributions exploited to fit the fail bits count/ECC capacity distribution on all the CWs measured after read disturb post-retention stress.

points left for satisfactory determination of the GPD parameters. Additionally, in [33] it is evidenced that the standard practice when choosing a threshold, is to select the lowest threshold possible for which the limit model (i.e., the GPD) provides a reasonable approximation for the exceedances.

A method to identify the correct threshold lies in the use of the mean residual life (MRL) plot combined with the stability plots of the GPD parameters. Concerning the former, the locus of points defined as:

$$\left\{\left(u, \frac{1}{n_u}\sum_{i=1}^{n_u}\left(x_{(i)} - u\right)\right) : u < x_{max}\right\} \tag{8}$$

where $x_{(1)}, \ldots, x_{(n_u)}$ are the n_u CWs exceeding the threshold u and x_{max} is the largest of the x_i, should be approximately linear in a ROI of u to define a proper threshold. For the latter, it is important to check whether the estimated GPD parameters are stable (i.e., constant) in the ROI, but after the following transformation of the GPD scale parameter [18]:

$$\sigma^* = \sigma - \xi u. \tag{9}$$

It is also important to check whether the threshold is meaningful for reliability investigations. To this extent, since all our fail bits count data are normalized with respect to the ECC capacity we decided to set the threshold $u = 1$. Every exceedance will therefore represent an unrecoverable CW and therefore a reliability concern in storage applications. Please note that in our study case we set the ECC capacity matching that offered by state-of-the-art correction engines. Retrospectively, we evaluated that such choice also grants a good number of exceedances where to apply the POT-GPD fit.

Fig. 8 shows the validation of the threshold choice procedure for read disturb post-retention data. Since this scenario represents a critical case for the reliability, surely more than the read disturb post-endurance as evidenced in the previous sections of the work, we will base all our investigations on this corner.

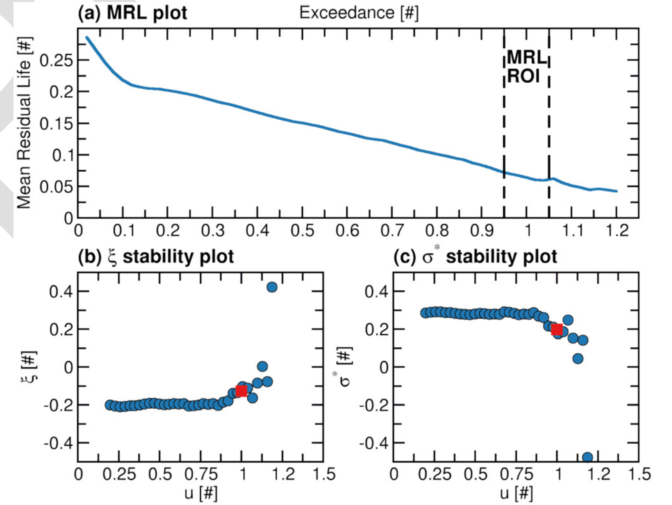


Fig. 8. (a) Mean residual life plot with a region of interest (ROI) highlighted. (b) and (c) Stability plots of GPD parameters. The CWs data are from post-retention read disturb tests [17].

C. Extending the EVA With POT-Weibull

The POT approach can be complemented with any probability distribution that can embed the concept of threshold. The GPD has been proven as one of the best statistical tools that fits all the modeling excesses problems, although this is not the only one. The Weibull distribution can be another viable approach, but not in the form of eq. (4). Indeed, the threshold concept must be included as in the following:

$$f(y) = \frac{\beta}{\alpha}\left(\frac{y-u}{\alpha}\right)^{\beta-1} e^{-\left(\frac{y-u}{\alpha}\right)^{\beta}} \tag{10}$$

where α and β are the same parameters defined in eq. (4), and u is the threshold defined in the previous section. This distribution is also referred as a three-parameters Weibull [34] that we will use as a benchmark for the GPD.

IEEE Proof

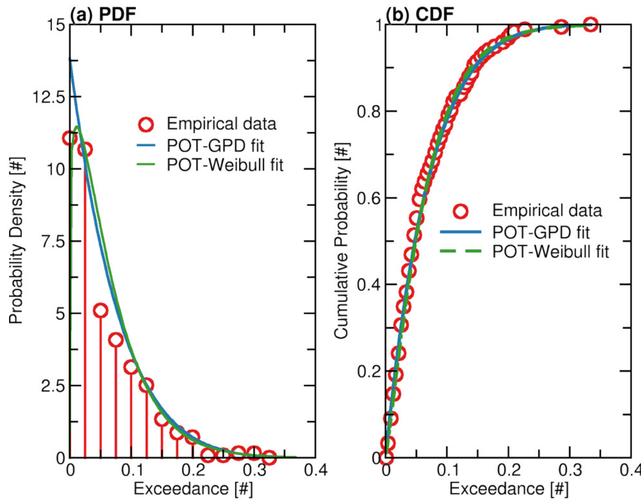


Fig. 9. (a) PDF and (b) CDF of the POT-GPD and POT-Weibull distributions devised in the modeling of the read disturb post-retention CW exceedances over threshold.

TABLE I
POT-GPD AND POT-WEIBULL PARAMETERS ESTIMATE ON READ DISTURB POST-RETENTION DATA USING THRESHOLD $u = 1$

	POT-GPD	POT-Weibull
$\hat{\xi}$	-0.13	-
$\hat{\sigma}$	0.07	-
$\hat{\alpha}$	-	0.07
$\hat{\beta}$	-	1.15
p-value (χ^2-test)	0.29	0.21

VI. ESTIMATING DIE-LEVEL RELIABILITY

A. Fitting Process and Model Cross-Validation

After the threshold choice process for read disturb post-retention data we evaluated the capability of the POT-GPD and POT-Weibull models to fit the exceedances of the CWs error distribution. In Fig. 9, we demonstrate that both the probability density function (PDF) and the CDF obtained through maximum likelihood estimation (MLE) well-fit the experimental data. Both models nicely describes the data. The estimated GPD parameters $\hat{\xi}$ and $\hat{\sigma}$ and the Weibull $\hat{\alpha}$ and $\hat{\beta}$ are reported in Table I. We run a goodness-of-fit χ^2 test with 0.05 confidence level to prove that the exceedances can be described with both POT approaches. The test passed with a p-value = 0.29 for POT-GPD and with a p-value = 0.21 for POT-Weibull (the higher the p-value the better it is), so there is no evidence to discard this statistical hypothesis.

B. Calculating the POT Return Level

We put the POT models at work to predict the die-level reliability of a 3D NAND Flash chip in the read disturb post-retention context starting with a limited number of blocks measured by our experimental setup. The goal of this process can be helpful as an example for system designers that requires a fast evaluation of the technological capabilities of the memory under test without requiring many empirical measurements.

The POT-GPD method enables such reliability assessment through the return level evaluation [31], [33]. The return level and the return period are two important concepts in the POT theory, thus requiring proper introduction. If we define a return period N of a CW that is measured in quantity of 3D NAND Flash memory blocks, the return level, x, is the threshold that is exceeded in one memory block with probability $\frac{1}{N}$. This is equivalent to claim that the return level x is exceeded on average once in N blocks. As an example, a CW with a fail bits count/ECC capacity ratio equal to 1.1 has a return period of 3 blocks if and only if the probability of observing a CW whose fail bits count/ECC capacity ratio higher than 1.1 in a block is $\frac{1}{3}$. In the POT theory, the return period is calculated as:

$$N = \underbrace{\underbrace{\frac{\text{number of CWs exceeding the threshold}}{\text{total number of CWs measured in blocks}}}_{\text{average number of exceedances per block}} \times m}_{\text{expected number of exceedances in m blocks}} \tag{11}$$

From the previous equation, it follows that N is the number of events over threshold between the occurrence of two consecutive CWs, both with a return period of m blocks. Hence, $\frac{1}{N}$ (i.e., the return level) is the probability of observing a CW with a return period of m blocks in one block. If we choose the CDF $F(x)$ of a specified probability distribution (i.e., the GPD or the Weibull used in the POT method) and $F(x) = 1 - \frac{1}{N}$, then $F(x)$ is the probability of observing any CW with a fail bits count/ECC capacity ratio less than or equal to x in one block.

Starting from that, we assumed that a 3D NAND Flash die is composed by 3000 blocks and then we estimated the return level per block in the case of the GPD distribution as:

$$x_m = u + \frac{\hat{\sigma}}{\hat{\xi}}\left[\left(m\hat{\zeta_u}\right)^{\hat{\xi}} - 1\right] \tag{12}$$

where m is the block number, $\hat{\zeta_u}$ is the probability to have an exceedance when a threshold u is considered, and $\hat{\sigma}$ and $\hat{\xi}$ are the estimated parameters of the GPD distribution. In the case of a Weibull distribution the previous return level equation becomes:

$$x_m = u + \hat{\alpha}\left[log\left(m\hat{\zeta_u}\right)^{1/\hat{\beta}}\right] \tag{13}$$

where $\hat{\alpha}$ and $\hat{\beta}$ are the estimated parameters of the three-parameters Weibull distribution, respectively.

The results in Fig. 10 for read disturb post-retention measurements evidence the return level to be expected for 3000 blocks also considering the 95% confidence interval for the GPD and Weibull distributions parameters estimates. From the return level analysis, we infer two results: i) the empirical data falls out of the POT-Weibull return level confidence interval for some points; ii) for a high number of blocks, the POT-GPD provides an optimistic estimation of the return level with respect to the POT-Weibull (lower fail bits count/ECC capacity ratio) while providing a larger return level confidence interval. We also run a cross-validation test using a Holdout methodology where 70% of the CWs tested in the experiments are used for training the statistical models and the remainder 30% are used for testing the POT models. On a total of 1000 cross-validation splits we report that the median p-value of the POT-GPD approach is slightly higher than that of the POT-Weibull, justifying the better prediction capabilities of the former model.

IEEE Proof

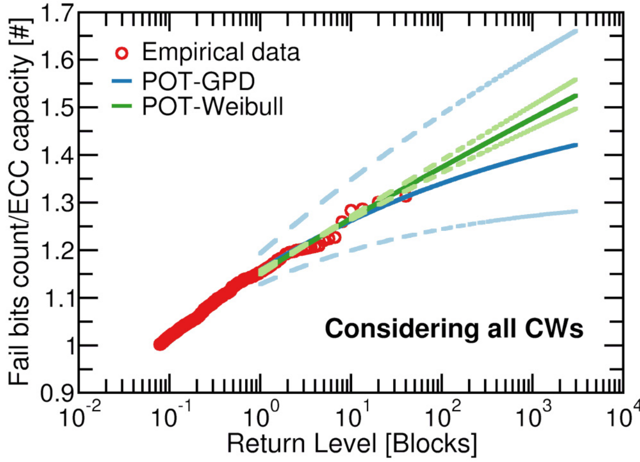


Fig. 10. Return level estimate for read disturb post-endurance of the POT-GPD and the POT-Weibull model with 95% confidence interval.

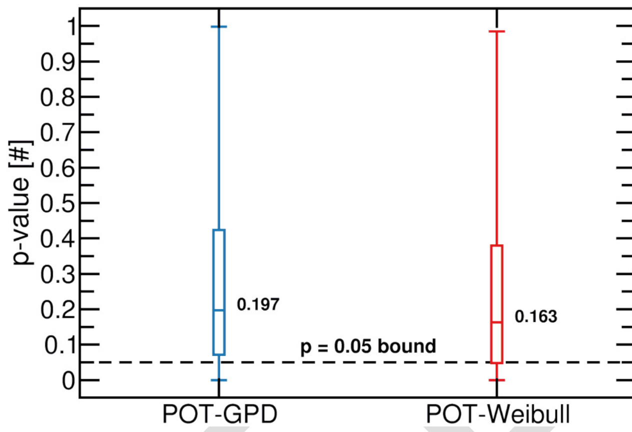


Fig. 11. Boxplot of the cross-validation splits performed for POT-GPD and POT-Weibull methods.

All these results clearly indicate that in mass storage application like SSDs and MMCs, where many blocks are considered for high storage capacity, advanced protection concerning the read disturb post-retention must be ensured since it is highly probable to encounter an unfavorable situation (i.e., unrecoverable errors) in some of the blocks constituting the memory die. This requires additional effort at system level to mitigate the 3D NAND Flash error probability.

C. Bootstrapping the POT Estimates

Since the POT-GPD and the POT-Weibull parameters are obtained through an MLE process we had to retrieve their confidence interval through a bootstrap analysis of the parameters with 1000 replica of the exceedances' dataset. A single bootstrap replica is a random sample of size n_u defined as $(x_1^*, x_2^*, \ldots, x_{nu}^*)$ drawn with replacement from the exceedances population of n_u samples retrieved with the procedure described in the former section of this work. In this case, the bootstrap data set consists of members of the original data set, some appearing zero times, some appearing once or multiple times. Fig. 12 shows a quantile-quantile plot proving a normal distribution of the POT-GPD parameters on which

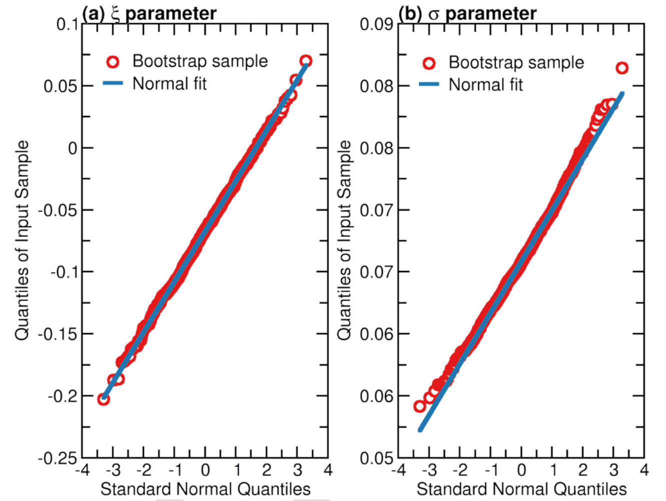


Fig. 12. Bootstrap simulation on the GPD parameters by resampling 1000 times the exceedances CW in the read disturb post-retention dataset [17].

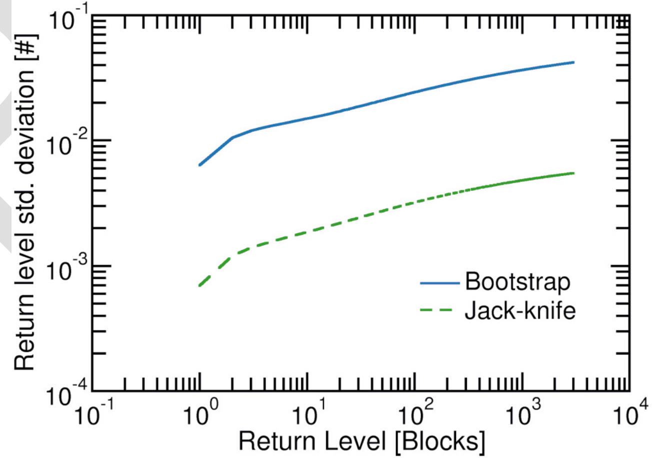


Fig. 13. Standard deviation in return level estimates depending on the chosen resampling technique. Solid lines are read disturb post-endurance data whereas dashed lines are post-retention.

it is easy to extract the confidence interval. Similar results (not shown) are achieved for the POT-Weibull. Nevertheless, we must report that this procedure has some issues in the lower quantiles of the normal distribution. This is ascribed to the MLE process convergence to a boundary point of the parameters space for some bootstrap samples.

Finally, we tried another resampling technique to check if we would achieve consistent results in the POT-GPD and POT-Weibull parameters estimation, namely the Jack-knife resampling. In this technique, if the original dataset of n_u exceedances is employed, the $i-th$ jack-knife sample is defined as:

$$x_{(i)} = (x_1; \ldots; x_{i-1}; x_{i+1}; \ldots; x_{nu}) \quad i = 1; \ldots; n_u. \quad (14)$$

A calculation of this method has been performed with a commercial tool for matrix data manipulation. To compare the prediction accuracy for both resampling technique we plotted the standard deviation of the return level predictions as a function of the return level calculated with (12). As we can see in Fig. 13, the jack-knife resampling provides the smallest

460 standard deviation for estimates (there is a difference up to 461 40 times at die level prediction) performed for read disturb 462 post-retention. This result is attributed to a small variation of 463 the new generated samples (the replica datasets differ for a 464 single value). To this extent, Jack-knife resampling technique 465 is not well suited to be used together with the POT approach, 466 since generated replicas are not so different, hence, estima467 tions based on these samples differ slightly and could lead to 468 optimistic predictions.

469 VII. Conclusion

470 In this work, we validated the POT methodology as a tech471 nique for EVA to be applied on a study case like the read 472 disturb reliability modeling in 3D NAND Flash memories. 473 The effectiveness of the model proven its applicability in 474 die level reliability predictions of the number of errors per 475 CW in an important scenario like the post-retention use case. 476 The methodology could be beneficial for storage system level 477 designers dealing with error mitigation schemes. In future, we 478 plan to apply the methodology to consider other 3D NAND 479 Flash reliability threats and to model extreme events in SSD 480 platforms studied at architectural level.

481 References

482 [1] R. Micheloni, S. Aritome, and L. Crippa, "Array architectures for 3-D 483 NAND flash memories," *Proc. IEEE*, vol. 105, no. 9, pp. 1634–1649, 484 Sep. 2017, doi: 10.1109/JPROC.2017.2697000.

485 [2] A. S. Spinelli, C. M. Compagnoni, and A. L. Lacaita, "Reliability of 486 NAND flash memories: Planar cells and emerging issues in 3D devices," 487 *Computers*, vol. 6, no. 2, p. 16, 2017, doi: 10.3390/computers6020016.

488 [3] T. A. Marquart, "Solid-state-drive qualification and reliability strategy," 489 in *Proc. IEEE Int. Integr. Rel. Workshop (IIRW)*, South Lake Tahoe, CA, 490 USA, Oct. 2015, pp. 3–6, doi: 10.1109/IIRW.2015.7437056.

491 [4] N. Mielke *et al.*, "Bit error rate in NAND flash memories," in *Proc.* 492 *IEEE Int. Rel. Phys. Symp.*, Phoenix, AZ, USA, Apr. 2008, pp. 9–19, 493 doi: 10.1109/RELPHY.2008.4558857.

494 [5] T. Parnell, N. Papandreou, T. Mittelholzer, and H. Pozidis, "Modelling of 495 the threshold voltage distributions of sub-20nm NAND flash memory," 496 in *Proc. IEEE Global Commun. Conf.*, Austin, TX, USA, Dec. 2014, 497 pp. 2351–2356, doi: 10.1109/GLOCOM.2014.7037159.

498 [6] N.-J. Wang *et al.*, "Statistical analysis of bit-errors distribution 499 for reliability of 3-D NAND flash memories," in *Proc. IEEE Int.* 500 *Rel. Phys. Symp. (IRPS)*, Dallas, TX, USA, Apr. 2020, pp. 1–5, 501 doi: 10.1109/IRPS45951.2020.9128993.

502 [7] C. Zambelli, R. Micheloni, and P. Olivo, "Reliability chal503 lenges in 3D NAND flash memories," in *Proc. IEEE 11th Int.* 504 *Memory Workshop (IMW)*, Monterey, CA, USA, May 2019, pp. 1–4, 505 doi: 10.1109/IMW.2019.8739741.

506 [8] L. Zuolo, C. Zambelli, R. Micheloni, and P. Olivo, "Solid507 state drives: Memory driven design methodologies for optimal 508 performance," *Proc. IEEE*, vol. 105, no. 9, pp. 1589–1608, Sep. 2017, 509 doi: 10.1109/JPROC.2017.2733621.

510 [9] T. Zhang, *Using LDPC Codes in SSD—Challenges and Solutions*, Flash 511 Memory Summit, Santa Clara, CA, USA, Aug. 2012.

512 [10] E. F. Haratsch, *LDPC Code Concepts and Performance on High-* 513 *Density Flash Memory*, Flash Memory Summit, Santa Clara, CA, USA, 514 Aug. 2014.

515 [11] N. R. Mielke, R. E. Frickey, I. Kalastirsky, M. Quan, D. Ustinov, 516 and V. J. Vasudevan, "Reliability of solid-state drives based on NAND 517 flash memory," *Proc. IEEE*, vol. 105, no. 9, pp. 1725–1750, Sep. 2017, 518 doi: 10.1109/JPROC.2017.2725738.

519 [12] S. Im and D. Shin, "Flash-aware RAID techniques for dependable and 520 high-performance flash memory SSD," *IEEE Trans. Comput.*, vol. 60, 521 no. 1, pp. 80–92, Jan. 2011, doi: 10.1109/TC.2010.197.

522 [13] J. Kim, E. Lee, J. Choi, D. Lee, and S. H. Noh, "Chip-level RAID 523 with flexible stripe size and parity placement for enhanced SSD relia524 bility," *IEEE Trans. Comput.*, vol. 65, no. 4, pp. 1116–1130, Apr. 2016, 525 doi: 10.1109/TC.2014.2375179.

[14] Y. Li, P. P. C. Lee, and J. C. S. Lui, "Analysis of reliability dynamics 526 of SSD RAID," *IEEE Trans. Comput.*, vol. 65, no. 4, pp. 1131–1144, 527 Apr. 2016, doi: 10.1109/TC.2014.2349505. 528

[15] C. Zambelli, A. Marelli, R. Micheloni, and P. Olivo, "Modeling the 529 endurance reliability of intradisk RAID solutions for mid-1X TLC 530 NAND flash solid-state drives," *IEEE Trans. Device Mater. Rel.*, vol. 17, 531 no. 4, pp. 713–721, Dec. 2017, doi: 10.1109/TDMR.2017.2749639. 532

[16] A. Grossi, L. Zuolo, F. Restuccia, C. Zambelli, and P. Olivo, 533 "Quality-of-service implications of enhanced program algorithms for 534 charge-trapping NAND in future solid-state drives," *IEEE Trans.* 535 *Device Mater. Rel.*, vol. 15, no. 3, pp. 363–369, Sep. 2015, 536 doi: 10.1109/TDMR.2015.2448108. 537

[17] C. Zambelli, L. Crippa, R. Micheloni, and P. Olivo, "Points-over- 538 threshold statistics for post-retention read disturb reliability in 3D NAND 539 flash," in *Proc. IEEE Int. Integr. Rel. Workshop (IIRW)*, South Lake 540 Tahoe, CA, USA, 2020, pp. 1–5. 541

[18] M. R. Leadbetter, "On a basis for 'peaks over threshold' 542 modeling," *Stat. Probab. Lett.*, vol. 12, no. 4, pp. 357–362, 1991, 543 doi: 10.1016/0167-7152(91)90107-3. 544

[19] K. Ha, J. Jeong, and J. Kim, "A read-disturb management technique for 545 high-density NAND flash memory," in *Proc. 4th Asia–Pac. Workshop* 546 *Syst.*, 2013, pp. 1–6, doi: 10.1145/2500727.2500743. 547

[20] C. Zambelli *et al.*, "Characterization of TLC 3D-NAND flash endurance 548 through machine learning for LDPC code rate optimization," in *Proc.* 549 *IEEE Int. Memory Workshop (IMW)*, Monterey, CA, USA, 2017, 550 pp. 1–4, doi: 10.1109/IMW.2017.7939074. 551

[21] *Electrically Erasable Programmable Rom (EEPROM) Program/Erase* 552 *Endurance and Data Retention Test* document JESD22-A117, JEDEC, 553 Arlington, VA, USA, Oct. 2018. 554

[22] C. Zambelli, P. Olivo, L. Crippa, A. Marelli, and R. Micheloni, 555 "Uniform and concentrated read disturb effects in mid-1X TLC NAND 556 flash memories for enterprise solid state drives," in *Proc. IEEE* 557 *Int. Rel. Phys. Symp. (IRPS)*, Monterey, CA, USA, 2017, pp. 1–4, 558 doi: 10.1109/IRPS.2017.7936387. 559

[23] C. Zambelli, R. Micheloni, S. Scommegna, and P. Olivo, "First evidence 560 of temporary read errors in TLC 3D-NAND flash memories exiting from 561 an idle state," *IEEE J. Electron Devices Soc.*, vol. 8, pp. 99–104, 2020, 562 doi: 10.1109/JEDS.2020.2965648. 563

[24] (2019). *Microsemi PM8609 NVMe2032 Flashtec NVMe* 564 *Controller*. [Online]. Available: https://www.microsemi.com/product- 565 directory/storage-ics/3687-flashtec-nvme-controllers 566

[25] N. Papandreou *et al.*, "Characterization and analysis of bit errors in 3D 567 TLC NAND flash memory," in *Proc. IEEE Int. Rel. Phys. Symp. (IRPS)*, 568 Monterey, CA, USA, 2019, pp. 1–6, doi: 10.1109/IRPS.2019.8720454. 569

[26] F. Wang *et al.*, "Lateral charge migration induced abnormal read disturb 570 in 3D charge-trapping NAND flash memory," *Appl. Phys. Exp.*, vol. 13, 571 no. 5, Apr. 2020, Art. no. 054002, doi: 10.35848/1882-0786/ab8729. 572

[27] Y. Cai, Y. Luo, S. Ghose, and O. Mutlu, "Read disturb errors in MLC 573 NAND flash memory: Characterization, mitigation, and recovery," in 574 *Proc. IEEE/IFIP Int. Conf. Depend. Syst. Netw.*, Rio de Janeiro, Brazil, 575 Jun. 2015, pp. 438–449, doi: 10.1109/DSN.2015.49. 576

[28] V. Taranalli, H. Uchikawa, and P. H. Siegel, "On the capacity of the 577 beta-binomial channel model for multi-level cell flash memories," *IEEE* 578 *J. Sel. Areas Commun.*, vol. 34, no. 9, pp. 2312–2324, Sep. 2016, 579 doi: 10.1109/JSAC.2016.2603660. 580

[29] Y. Luo, S. Ghose, Y. Cai, E. F. Haratsch, and O. Mutlu, 581 "Improving 3D NAND flash memory lifetime by tolerating early 582 retention loss and process variation," 2018. [Online]. Available: 583 arXiv:1807.05140. 584

[30] Y. Cai, E. F. Haratsch, O. Mutlu, and K. Mai, "Threshold voltage dis- 585 tribution in MLC NAND flash memory: Characterization, analysis, and 586 modeling," in *Proc. Design Autom. Test Eur. Conf. Exhibit. (DATE)*, 587 Grenoble, France, 2013, pp. 1285–1290, doi: 10.7873/DATE.2013.266. 588

[31] C. Stander, "Analysis of extreme events in the coastal engineer- 589 ing environment," M.S. thesis, Dept. Appl. Math. Stellenbosch Univ., 590 Stellenbosch, South Africa, Dec. 2015. 591

[32] N. Teena, V. S. Kumar, K. Sudheesh, and R. Sajeev, "Statistical anal- 592 ysis on extreme wave height," *Nat. Hazards J. Int. Soc. Prevent.* 593 *Mitigation Nat. Hazards*, vol. 64, no. 1, pp. 223–236, Oct. 2012, 594 doi: 10.1007/s11069-012-0229-y. 595

[33] S. Coles, *An Introduction to Statistical Modeling of Extreme Values*. 596 London, U.K.: Springer, 2001. 597 AQ3

[34] E. Merran, N. Hastings, and B. Peacock, *Statistical Distributions*, 598 2nd ed. New York, NY, USA: Wiley, 1993. 599

IEEE Proof

AUTHOR QUERIES

AUTHOR PLEASE ANSWER ALL QUERIES

PLEASE NOTE: We cannot accept new source files as corrections for your paper. If possible, please annotate the PDF proof we have sent you with your corrections and upload it via the Author Gateway. Alternatively, you may send us your corrections in list format. You may also upload revised graphics via the Author Gateway.

Carefully check the page proofs (and coordinate with all authors); additional changes or updates WILL NOT be accepted after the article is published online/print in its final form. Please check author names and affiliations, funding, as well as the overall article for any errors prior to sending in your author proof corrections. Your article has been peer reviewed, accepted as final, and sent in to IEEE. No text changes have been made to the main part of the article as dictated by the editorial level of service for your publication.

AQ1: According to our records, Luca Crippa is listed as a Member, IEEE. Please verify.

AQ2: Please confirm or add details for any funding or financial support for the research of this article.

AQ3: Please confirm if the location and publisher information for Reference [33] is correct as set.

IEEE Proof

Investigating 3D NAND Flash Read Disturb Reliability With Extreme Value Analysis

AQ1 Cristian Zambelli, *Member, IEEE*, Luca Crippa, *Member, IEEE*, Rino Micheloni, *Senior Member, IEEE*, and Piero Olivo

***Abstract*—The storage systems relying on the 3D NAND Flash technology require an extensive modeling of their reliability in different working corners. This enables the deployment of system-level management routines that do not compromise the overall performance and reliability of the system itself. Dedicated parametric statistical models have been developed so far to capture the evolution of the memory reliability, although limiting the description to an average behavior rather than extreme cases that can disrupt the storage functionality. In this work, we validate the application of an extreme statistics tool, namely the Points-Over-Threshold method, to characterize the read disturb reliability of a 3D NAND Flash chip. Such technique proved that the die reliability characterized through extreme events analysis can be predicted using a low number of samples and generally holds good prediction features for distribution tail events.**

***Index Terms*—3D-TLC NAND flash, read disturb, reliability, points over threshold.**

I. Introduction

MODELLING the reliability of the 3D NAND Flash memory technology [1], [2] is still an important task to be performed as a support for storage system designers dedicated to Solid State Drives (SSDs) or Multi Media Card (MMC) products development. Indeed, all the firmware solutions implemented in their controllers (i.e., the computing core of SSDs and MMCs) whose goal is mitigating the inherent bit error rate (BER) exposed in different storage working conditions (e.g., endurance stress, data retention at high temperature, etc.) are well founded on memory reliability models [3]. To this extent, dedicated parametric statistical models [4]–[6] have been developed so far to capture the evolution of the memory's errors distribution through well-known statistical frameworks (defined as probability distributions) like Gaussian, Binomial, Poisson, Gamma, and so on. However, the large process-induced variability of the error characteristics in 3D NAND Flash devices [7] combined with an intrinsic difficulty in testing all the possible permutations of the memory working corners during lifetime and on a relevant statistical population, could hamper an accurate description of the distribution upper tail. It is worth to mention that on this part of the errors distribution, the storage system designers spend a significant effort to tailor the Error Correction Codes (ECCs) [8] strength and the secondary correction schemes like soft decoding [9], [10], Moving Read References [11], and even RAID [12]–[15]. Therefore, the more precise and accurate is the model the less is the probability to incur in storage performance slowdowns due to improperly calibrated error correction techniques [16].

In [17], we proposed for the first time to apply a parametric model commonly exploited in extreme value analysis (EVA) for natural sciences and econometrics to 3D NAND Flash errors distribution upper tail modeling, namely the Point Over Threshold (POT) method [18]. The work demonstrated the potential of this methodology combined with the Generalized Pareto Distribution (GPD) in the analysis of the read disturb stress after data retention. We picked that memory working condition since it is known to represent a critical use case in data center applications for Big Data analytics performed on cold data [19]. Starting from our previous work, we extended the study in twofold directions: the first related to the characterization and estimation of the read disturb also for hot data scenario (i.e., read after many updates) mimicked by an endurance stress, showing that it is not critical for the reliability as in retention conditions; the second concerning the cross-validation of the POT and the extension of its implementation with a three-parameters Weibull distribution. This will demonstrate, with a proper confidence, that POT is a powerful statistical tool to estimate the 3D NAND Flash die reliability.

II. Experimental Setup

A. Devices Under Test

The experimental activity in this work is based on the characterization of an off-the-shelf sub-100 layers 3D NAND Flash memory product implementing the Triple Level Cell (TLC) paradigm (see Fig. 1). Such technology is considered, based on its endurance and retention rating, as a mass storage medium for enterprise SSD applications. The statistical sample under investigation is composed by all the pages in every physical layer of 40 memory blocks distributed on multiple dies and chips to account for process-induced variability [7]. Since we are testing a TLC memory, we consider all the page types in the analysis (i.e., LSB-Least Significant Bit page, CSB-Center Significant Bit page, and MSB-Most Significant Bit page). Each page is sized 16 Kbytes plus the spare bytes exploited for

Manuscript received August 2, 2021; accepted August 27, 2021. *(Corresponding author: Cristian Zambelli.)*

Cristian Zambelli and Piero Olivo are with the Dipartimento di Ingegneria, Universita degli Studi di Ferrara, 44122 Ferrara, Italy (e-mail: cristian.zambelli@unife.it).

Luca Crippa and Rino Micheloni are with the Flash Signal Processing Labs, Microchip Corporation, 20871 Vimercate, Italy.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TDMR.2021.3108941.

Digital Object Identifier 10.1109/TDMR.2021.3108941

1530-4388 © 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

IEEE Proof

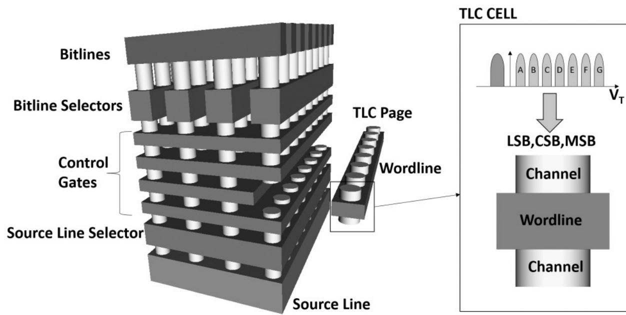


Fig. 1. TLC 3D NAND Flash architecture considered in this work [17].

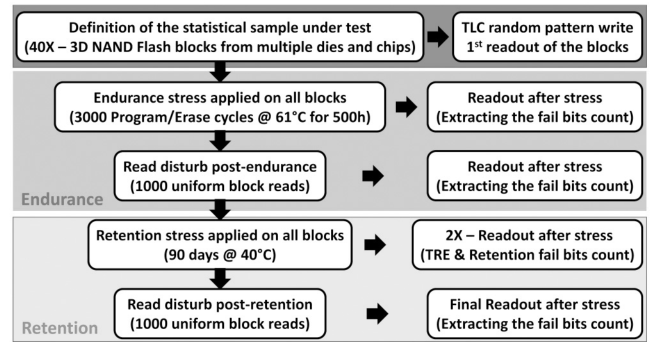


Fig. 2. Depiction of the test flow adopted in this work for the 3D NAND Flash reliability characterization.

data recovery purposes in case of data corruption. However, state-of-the-art ECCs implementations [8] work on subset of the page dimension, generally referred as a codeword (CW). In this work, the CW size is 4 Kbytes plus the spare bits, so that every page is constituted by 4 CWs. The statistical sample under test is constituted by 184320 CWs.

B. Test Flow for Reliability Assessments

The execution of the test flows required to extract the experimental data to fit with the statistical models presented in the paper is performed by the automated test equipment (ATE) presented in [20]. The system interfaces with the 3D NAND Flash chips at a 400 MT/s data rate and allows, for any applied test, to measure the number of corrupted bits (also known as fail bits count or errors number) in each CW. We remind that a bit is considered corrupted after reading from the 3D NAND Flash under test if its value changes from what has been previously written as a result of a reliability degradation process. Fig. 2 summarizes the test procedure performed. After the definition of the statistical sample under test, we write a random pattern on all the TLC pages (therefore on all the CWs) to rule out any topological dependency of the corrupted bits and perform a readout of the memory content. Then, the devices are submitted to an endurance stress up to the rated endurance of the technology (i.e., maximum number of sustainable block erase before unrecoverable errors) using a JEDEC-based cycling test [21]. The test consists in 3000 Program/Erase cycles at a 61 °C temperature for 500 hours. After the stress we perform a readout of all the CWs under test and we extract the number of corrupted bits for each CW. A read disturb is then performed post-endurance by employing a 1000 uniform block reads access pattern [22] on all tested blocks and performed another readout for fail bits count extraction. Immediately after the end of the read disturb post-endurance stress, a data retention stress is performed by placing all the devices under test in idle for 90 days at a 40 °C temperature. A double readout is performed at the end of the test to separate the Temporary Read Errors (TRE) effect typical of 3D NAND Flash architectures [23] from the retention stress results. Finally, an additional read disturb post-retention is performed and the fail bits count per CW are extracted accordingly.

III. 3D NAND Flash Read Disturb Characterization

We started the read disturb characterization on our devices by evaluating the Empirical Cumulative Distribution Function (ECDF) of the fail bits count extracted on all the CWs before and after the read disturb stress post-endurance and post-retention scenarios, as described in Fig. 2. For a given number of fail bits equal to t in a sample x, the ECDF is defined as the proportion of the values in $x \leq t$. We prefer to use the term ECDF rather than CDF since the latter one can be misleading as it is usually referenced to a theoretical probability distribution used to fit the experimental data, which is not our case here. Unfortunately, due to confidentiality reasons on the tested 3D NAND Flash samples we cannot disclose the ECDF of the fail bits count, but we have to normalize the number of corrupted bits on a CW to a defined entity. In this work, we normalized the fail bits count with respect to the ECC capacity offered by an advanced correction engine that incorporates secondary error correction schemes (e.g., read retry and soft-decoding) as well [9], [10], [24].

Fig. 3 shows that after endurance stress the read disturb generally increases the number of errors as expected from other studies in this context [25], [26]. This behavior is related to an over-programming of the memory cells that are not involved by the read operation due a moderate voltage applied to unselect them [27]. Looking at the median of the ECDFs per page type it is observed that the CWs belonging to MSB pages are those displaying the largest errors increase, even if there are some CWs in LSB pages (those in the associated ECDF tail) that are largely affected. However, the read disturb post-endurance is not particularly detrimental for the reliability since the ratio fail bits count/ECC capacity is well below one and displays a sufficient margin for safe operation without data corruption.

Fig. 4 shows the results of the same analysis replicated for the retention domain. In the errors analysis we also reported the ECDFs retrieved on the first readout of the memory blocks under test to evaluate the impact of the TRE, as discussed in the Section II of this work. We interestingly observe that the read disturb applied post-retention can recover part of the errors in a CW for all TLC page types. This has been explained in [26], by a charge redistribution mechanism occurring during

IEEE Proof

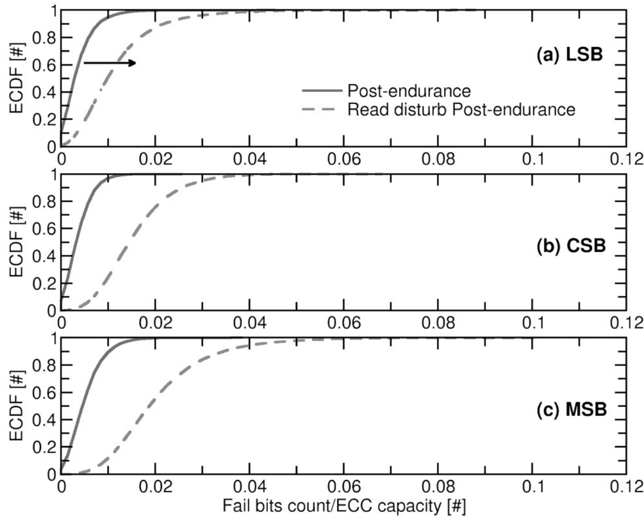


Fig. 3. ECDFs per TLC page type of the fail bits count normalized with respect to the ECC capacity in post-endurance stress and after the application of the read disturb. The ECDFs are extracted from all tested 3D NAND Flash CWs.

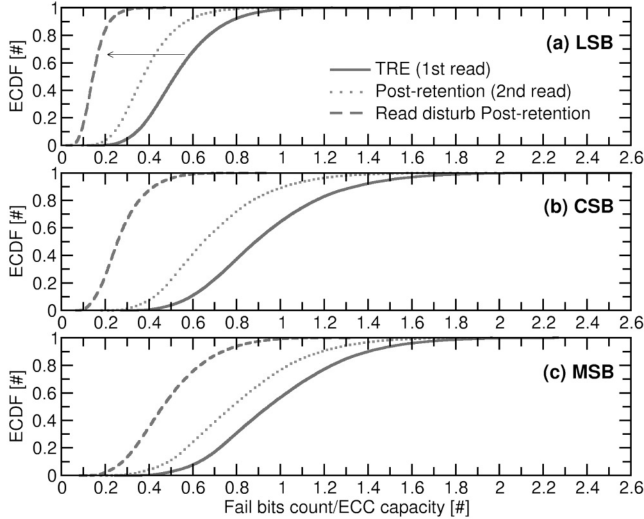


Fig. 4. ECDFs per TLC page type of the of the fail bits count normalized with respect to the ECC capacity in post-retention stress and after the application of the read disturb. The effect of the TRE is also evidenced in the figure.

the read operation. From the reliability standpoint, we note that the retention scenario is the most critical to address since the ratio fail bits count/ECC capacity can be greater than one (especially for MSB pages), thus hampering the data recovery operations.

Besides the errors' distribution characterization according to the TLC page type, we analyzed what is the contribution of the topological position (i.e., the layer position in a 3D NAND Flash block) on the fail bits count after the application of the read disturb. Fig. 5 shows the results of this investigation for the read disturb post-retention test case. We focus on this stress condition since it is the one triggering the highest number of errors during tests. We can note a large error variability among layers in a single 3D NAND Flash block after the post-retention read disturb and on top of this, there is a large error

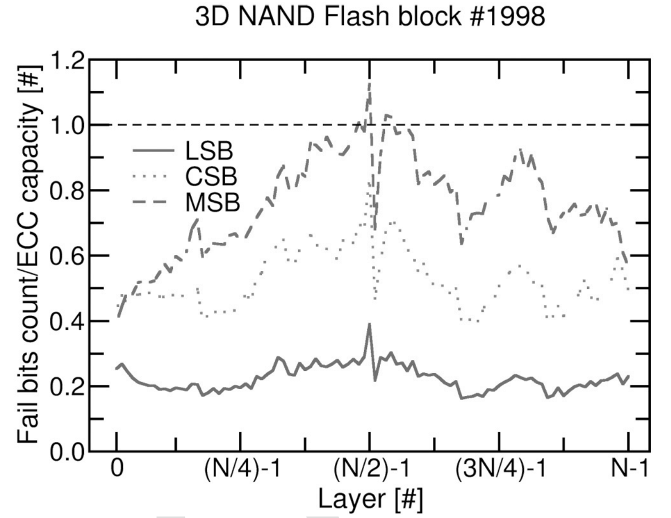


Fig. 5. Fail bits count normalized with respect to the ECC capacity characteristics per TLC page type as a function of layer position in a 3D NAND Flash block after read disturb post-retention. The ECC limit is highlighted for clarity with the black dashed line.

variation between LSB pages and CSB/MSB pages. This is a critical aspect that should be tackled by a statistical model developed to capture errors characteristics. Finally, to better understand the role of the read disturb on the errors count and therefore on the memory reliability we have calculated the error amplification (EA) factor as:

$$EA(i) = \frac{ERD(i)}{EPRE(i)} \tag{1}$$

where i is the layer position in the block from 0 to $N-1$ with N the number of layers exploited in the manufacturing of the 3D NAND Flash chip, *ERD* is the number of errors post-read disturb and *EPRE* the number of errors pre-read disturb. The EA is calculated both for endurance and retention test cases. Even if the endurance case is the one showing the largest EA, we must report once again that such scenario is not critical for the memory reliability since the errors amount remains always well below the ECC limits. On the contrary, even if the EA is below unity for post-retention read disturb, there are some critical conditions on which the errors count is higher than the ECC capacity and are worth to be modeled for future reliability considerations.

IV. Standard Statistical Modeling Approach

The common procedure adopted for modeling the fail bits count, and in general for all the parametric statistical frameworks applied to 3D NAND Flash data, is to fit the entire ECDF retrieved in a specific working condition. For our study case this is after a read disturb stress performed after endurance or after data retention at high temperature. The advantage of such parametric approach is to achieve a rapid estimation of the ECC capacity to cover errors and proven to be useful in many cases. The statistical modeling of CWs fail bits count distribution bases on the assumption that the corrupted bits in a CW are treated as independent events. By considering a CW length of n bytes, it is possible to calculate

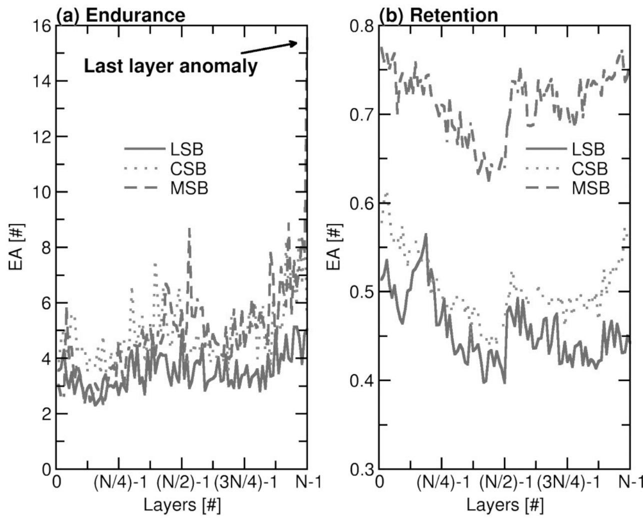


Fig. 6. EA factor for read disturb calculated as a function of the layer position in the memory block in endurance (a) and retention (b) test cases.

the probability of having k errors in the CW exhibiting a BER p using the binomial distribution probability density function as in [4]:

$$y = P_{error}(k|n, p) = \binom{n}{k} p^k (1-p)^{n-k} \tag{2}$$

However, considering that in 3D NAND Flash technology n (equal to 4 Kbytes plus spare bits in our work) is relatively larger than k and p is usually lower than 2×10^{-2}, the binomial approach starts to fail. Some alternative distributions like the beta-binomial [28], the Gamma [29], the Gamma-Poisson compound [6], or the Weibull [30] have been in consideration by the literature due to their capability in accounting the intrinsic variability of the memory technology. The easiest to calculate with software tools for numerical analysis are the Gamma and the Weibull distributions. The former is based on the following probability density function:

$$y = P_{error}(\lambda|\alpha, \beta) = \frac{\lambda^{\alpha-1}}{\Gamma(\alpha)\beta^{\alpha}} e^{-\left(\frac{\lambda}{\beta}\right)} \tag{3}$$

where λ is calculated as $n \cdot p$, α is the shape factor, and β is the scale factor of the distribution. The latter is:

$$y = P_{error}(k|\alpha, \beta) = \frac{\beta}{\alpha}\left(\frac{k}{\alpha}\right)^{\beta-1} e^{-\left(\frac{k}{\alpha}\right)^{\beta}} \tag{4}$$

with α and β being the shape and the scale factor of the distribution, respectively.

Unfortunately, due to the extreme variability characteristic of the fail bits retrieved in different locations of a 3D NAND Flash chip (see Fig. 5), both statical models (Gamma and Weibull) do not pass the χ^2 goodness-of-fit test (p-value = 0). One may still argue that even if the models do not pass the test, they are still valid for predictions of the ECDF distribution tails, whose practical applications are the selection of the ECC capacity to cover errors or the evaluation of the reliability margin. To this extent, we run a cross-validation test using a Holdout methodology where 70% of the CWs tested in the experiments are used for training the statistical models and the remainder 30% are used for testing its prediction accuracy. On a total of 1000 cross-validation splits, none of them passed the goodness-of-fit test. Analyzing the histogram count (see Fig. 7) of the fail bits count/ECC capacity and the resulting fits of the statistical models, we evidence an underestimation/overestimation of the empirical data distribution, possibly hampering the selection of the correct ECC capacity or its margin. The Gamma distribution performs better with respect to the Weibull, but still do not pass any statistical test. Same considerations can be drawn by looking at the Cumulative Distribution Function (CDF) modeled by the two statistical models. As in 3D NAND Flash reliability modeling we are mainly interested in the low probability tail of the errors distribution (i.e., extreme events), we need a framework capable to handle only that part of the empirical data.

V. The POT Method for Read Disturb EVA

A. Introducing the POT-GPD

Motivated by this, we explored a statistical framework related to EVA [18] that is commonly used in tail data analysis, namely the POT. By considering each 3D NAND Flash CW as a sequence of *i.i.d.* measurements $x_1, x_2, \ldots, x_n$, we can define as extreme events all the CWs that exceed a defined error threshold u for which we can define an exceedance as:

$$\{x_i : x_i \geq u\}. \tag{5}$$

If the exceedances are labeled as $x_{(1)}, \ldots, x_{(k)}$, it is possible to define a threshold excess as:

$$y_j = x_{(j)} - u \qquad j = 1, \ldots, k. \tag{6}$$

From the probability theory it is proven that a random variable Y_i based on the threshold excesses follows a GPD [18]. For a large enough threshold u, we can write its probability density function as:

$$f(y) = \sigma^{-1}\left(1 + \frac{\xi y}{\sigma}\right)^{-1-\xi^{-1}} \tag{7}$$

with the parameters ξ and σ being the distribution shape and scale factors, respectively.

B. Threshold Choice

The most critical operation in POT-GPD statistical modeling is the extrapolation of the best threshold to apply. The selection of an optimal threshold within a region of interest (ROI) requires a bias-variance trade-off and a knowledge of the ECC capabilities offered in the 3D NAND Flash data recovery processes. If the chosen model threshold is too low, the results are biased because of the model asymptotic assumption being invalid. In other words, a too low threshold will result in having exceedances not converging to the GPD, since the probability distribution is based on its capability of fitting extreme events [31]. On the other hand, if the threshold is too high, the variance is large due to few exceedances. In [32], it is stated that the threshold must be high enough for the exceedances over threshold to converge to the GPD, while the sample size should be large enough to ensure that there are enough data

IEEE Proof

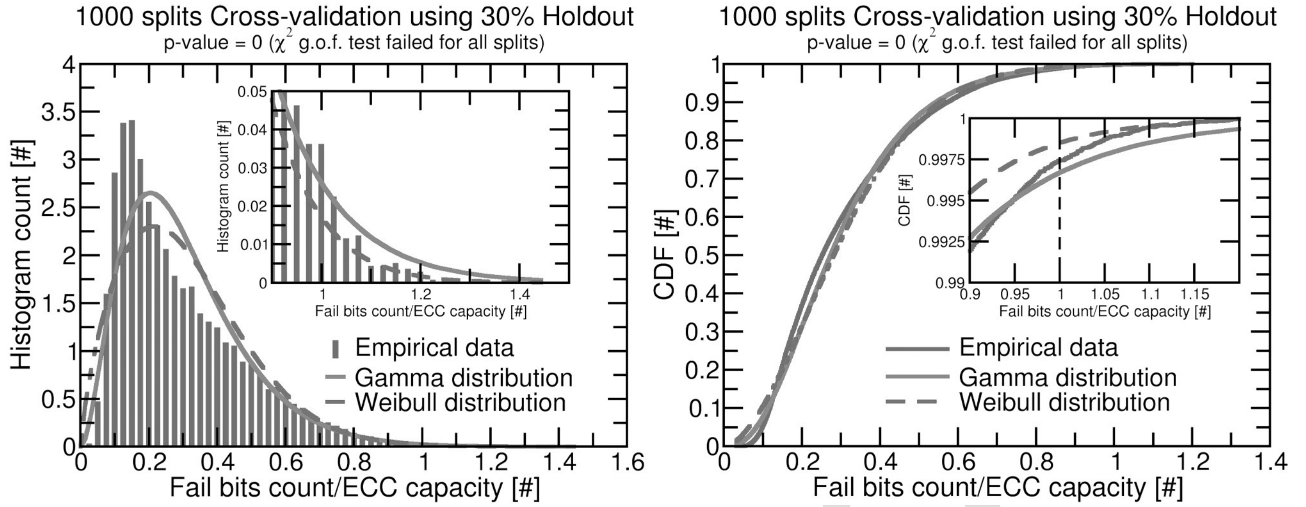


Fig. 7. Gamma and Weibull distributions exploited to fit the fail bits count/ECC capacity distribution on all the CWs measured after read disturb post-retention stress.

points left for satisfactory determination of the GPD parameters. Additionally, in [33] it is evidenced that the standard practice when choosing a threshold, is to select the lowest threshold possible for which the limit model (i.e., the GPD) provides a reasonable approximation for the exceedances.

A method to identify the correct threshold lies in the use of the mean residual life (MRL) plot combined with the stability plots of the GPD parameters. Concerning the former, the locus of points defined as:

$$\left\{\left(u, \frac{1}{n_u}\sum_{i=1}^{n_u}\left(x_{(i)} - u\right)\right) : u < x_{max}\right\} \tag{8}$$

where $x_{(1)}, \ldots, x_{(n_u)}$ are the n_u CWs exceeding the threshold u and x_{max} is the largest of the x_i, should be approximately linear in a ROI of u to define a proper threshold. For the latter, it is important to check whether the estimated GPD parameters are stable (i.e., constant) in the ROI, but after the following transformation of the GPD scale parameter [18]:

$$\sigma^* = \sigma - \xi u. \tag{9}$$

It is also important to check whether the threshold is meaningful for reliability investigations. To this extent, since all our fail bits count data are normalized with respect to the ECC capacity we decided to set the threshold $u = 1$. Every exceedance will therefore represent an unrecoverable CW and therefore a reliability concern in storage applications. Please note that in our study case we set the ECC capacity matching that offered by state-of-the-art correction engines. Retrospectively, we evaluated that such choice also grants a good number of exceedances where to apply the POT-GPD fit.

Fig. 8 shows the validation of the threshold choice procedure for read disturb post-retention data. Since this scenario represents a critical case for the reliability, surely more than the read disturb post-endurance as evidenced in the previous sections of the work, we will base all our investigations on this corner.

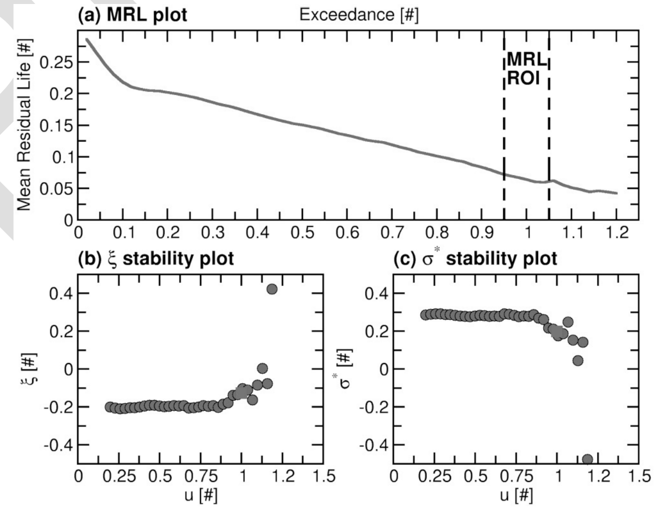


Fig. 8. (a) Mean residual life plot with a region of interest (ROI) highlighted. (b) and (c) Stability plots of GPD parameters. The CWs data are from post-retention read disturb tests [17].

C. Extending the EVA With POT-Weibull

The POT approach can be complemented with any probability distribution that can embed the concept of threshold. The GPD has been proven as one of the best statistical tools that fits all the modeling excesses problems, although this is not the only one. The Weibull distribution can be another viable approach, but not in the form of eq. (4). Indeed, the threshold concept must be included as in the following:

$$f(y) = \frac{\beta}{\alpha}\left(\frac{y-u}{\alpha}\right)^{\beta-1} e^{-\left(\frac{y-u}{\alpha}\right)^{\beta}} \tag{10}$$

where α and β are the same parameters defined in eq. (4), and u is the threshold defined in the previous section. This distribution is also referred as a three-parameters Weibull [34] that we will use as a benchmark for the GPD.

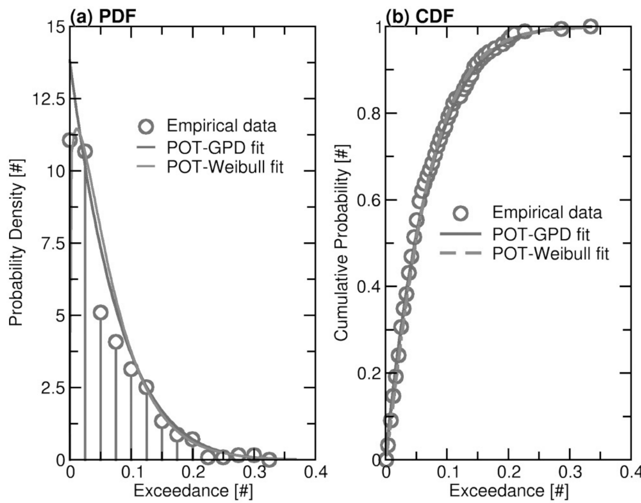


Fig. 9. (a) PDF and (b) CDF of the POT-GPD and POT-Weibull distributions devised in the modeling of the read disturb post-retention CW exceedances over threshold.

TABLE I
POT-GPD AND POT-WEIBULL PARAMETERS ESTIMATE ON READ DISTURB POST-RETENTION DATA USING THRESHOLD $u = 1$

	POT-GPD	POT-Weibull
$\hat{\xi}$	-0.13	-
$\hat{\sigma}$	0.07	-
$\hat{\alpha}$	-	0.07
$\hat{\beta}$	-	1.15
p-value (χ^2-test)	0.29	0.21

341 ## VI. ESTIMATING DIE-LEVEL RELIABILITY

342 ### A. Fitting Process and Model Cross-Validation

343 After the threshold choice process for read disturb post-
344 retention data we evaluated the capability of the POT-GPD
345 and POT-Weibull models to fit the exceedances of the CWs
346 error distribution. In Fig. 9, we demonstrate that both the
347 probability density function (PDF) and the CDF obtained
348 through maximum likelihood estimation (MLE) well-fit the
349 experimental data. Both models nicely describes the data. The
350 estimated GPD parameters $\hat{\xi}$ and $\hat{\sigma}$ and the Weibull $\hat{\alpha}$ and
351 $\hat{\beta}$ are reported in Table I. We run a goodness-of-fit χ^2 test
352 with 0.05 confidence level to prove that the exceedances can
353 be described with both POT approaches. The test passed with
354 a p-value = 0.29 for POT-GPD and with a p-value = 0.21 for
355 POT-Weibull (the higher the p-value the better it is), so there
356 is no evidence to discard this statistical hypothesis.

357 ### B. Calculating the POT Return Level

358 We put the POT models at work to predict the die-level
359 reliability of a 3D NAND Flash chip in the read disturb post-
360 retention context starting with a limited number of blocks
361 measured by our experimental setup. The goal of this pro-
362 cess can be helpful as an example for system designers that
363 requires a fast evaluation of the technological capabilities
364 of the memory under test without requiring many empirical
365 measurements.

366 The POT-GPD method enables such reliability assessment
367 through the return level evaluation [31], [33]. The return level
368 and the return period are two important concepts in the POT
369 theory, thus requiring proper introduction. If we define a return
period N of a CW that is measured in quantity of 3D NAND 370
Flash memory blocks, the return level, x, is the threshold that 371
is exceeded in one memory block with probability $\frac{1}{N}$. This 372
is equivalent to claim that the return level x is exceeded on 373
average once in N blocks. As an example, a CW with a fail bits 374
count/ECC capacity ratio equal to 1.1 has a return period of 3 375
blocks if and only if the probability of observing a CW whose 376
fail bits count/ECC capacity ratio higher than 1.1 in a block 377
is $\frac{1}{3}$. In the POT theory, the return period is calculated as: 378

$$N = \underbrace{\underbrace{\frac{\text{number of CWs exceeding the threshold}}{\text{total number of CWs measured in blocks}}}_{\text{average number of exceedances per block}} \times m}_{\text{expected number of exceedances in m blocks}} \quad (11)$$ 379

From the previous equation, it follows that N is the number 380
of events over threshold between the occurrence of two con- 381
secutive CWs, both with a return period of m blocks. Hence, 382
$\frac{1}{N}$ (i.e., the return level) is the probability of observing a CW 383
with a return period of m blocks in one block. If we choose the 384
CDF $F(x)$ of a specified probability distribution (i.e., the GPD 385
or the Weibull used in the POT method) and $F(x) = 1 - \frac{1}{N}$, 386
then $F(x)$ is the probability of observing any CW with a fail 387
bits count/ECC capacity ratio less than or equal to x in one 388
block. 389

Starting from that, we assumed that a 3D NAND Flash die 390
is composed by 3000 blocks and then we estimated the return 391
level per block in the case of the GPD distribution as: 392

$$x_m = u + \frac{\hat{\sigma}}{\hat{\xi}}\left[\left(m\hat{\zeta_u}\right)^{\hat{\xi}} - 1\right] \quad (12)$$ 393

where m is the block number, $\hat{\zeta_u}$ is the probability to have an 394
exceedance when a threshold u is considered, and $\hat{\sigma}$ and $\hat{\xi}$ are 395
the estimated parameters of the GPD distribution. In the case 396
of a Weibull distribution the previous return level equation 397
becomes: 398

$$x_m = u + \hat{\alpha}\left[log\left(m\hat{\zeta_u}\right)^{1/\hat{\beta}}\right] \quad (13)$$ 399

where $\hat{\alpha}$ and $\hat{\beta}$ are the estimated parameters of the three- 400
parameters Weibull distribution, respectively. 401

The results in Fig. 10 for read disturb post-retention mea- 402
surements evidence the return level to be expected for 3000 403
blocks also considering the 95% confidence interval for the 404
GPD and Weibull distributions parameters estimates. From 405
the return level analysis, we infer two results: i) the empir- 406
ical data falls out of the POT-Weibull return level confidence 407
interval for some points; ii) for a high number of blocks, the 408
POT-GPD provides an optimistic estimation of the return level 409
with respect to the POT-Weibull (lower fail bits count/ECC 410
capacity ratio) while providing a larger return level confidence 411
interval. We also run a cross-validation test using a Holdout 412
methodology where 70% of the CWs tested in the experiments 413
are used for training the statistical models and the remain- 414
der 30% are used for testing the POT models. On a total of 415
1000 cross-validation splits we report that the median p-value 416
of the POT-GPD approach is slightly higher than that of the 417
POT-Weibull, justifying the better prediction capabilities of the 418
former model. 419

IEEE Proof

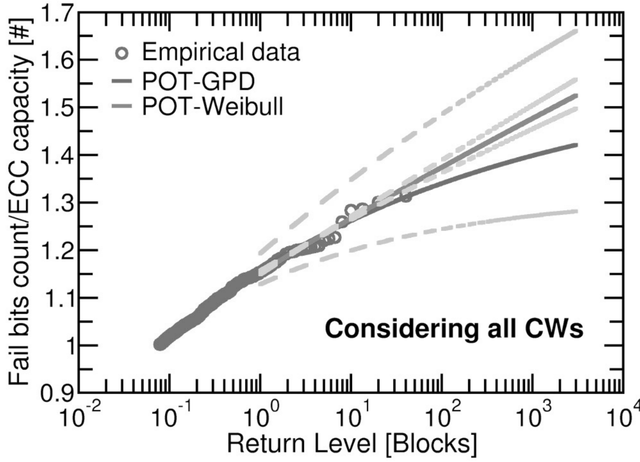


Fig. 10. Return level estimate for read disturb post-endurance of the POT-GPD and the POT-Weibull model with 95% confidence interval.

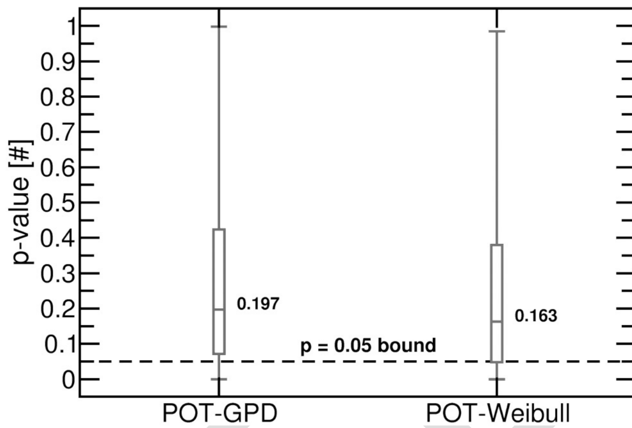


Fig. 11. Boxplot of the cross-validation splits performed for POT-GPD and POT-Weibull methods.

All these results clearly indicate that in mass storage application like SSDs and MMCs, where many blocks are considered for high storage capacity, advanced protection concerning the read disturb post-retention must be ensured since it is highly probable to encounter an unfavorable situation (i.e., unrecoverable errors) in some of the blocks constituting the memory die. This requires additional effort at system level to mitigate the 3D NAND Flash error probability.

C. Bootstrapping the POT Estimates

Since the POT-GPD and the POT-Weibull parameters are obtained through an MLE process we had to retrieve their confidence interval through a bootstrap analysis of the parameters with 1000 replica of the exceedances' dataset. A single bootstrap replica is a random sample of size n_u defined as $(x_1^*, x_2^*, \ldots, x_{nu}^*)$ drawn with replacement from the exceedances population of n_u samples retrieved with the procedure described in the former section of this work. In this case, the bootstrap data set consists of members of the original data set, some appearing zero times, some appearing once or multiple times. Fig. 12 shows a quantile-quantile plot proving a normal distribution of the POT-GPD parameters on which

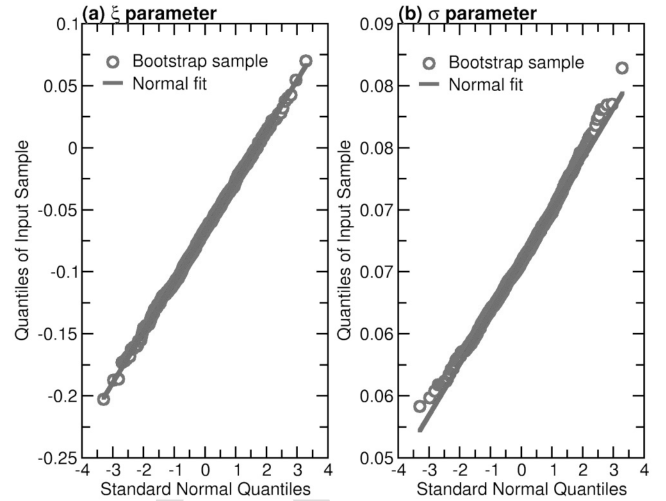


Fig. 12. Bootstrap simulation on the GPD parameters by resampling 1000 times the exceedances CW in the read disturb post-retention dataset [17].

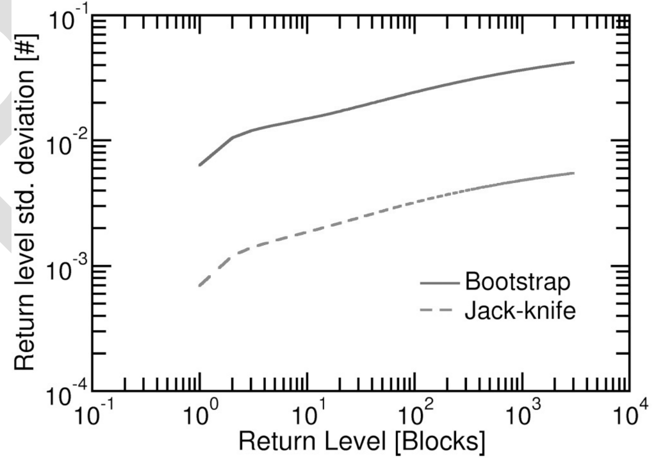


Fig. 13. Standard deviation in return level estimates depending on the chosen resampling technique. Solid lines are read disturb post-endurance data whereas dashed lines are post-retention.

it is easy to extract the confidence interval. Similar results (not shown) are achieved for the POT-Weibull. Nevertheless, we must report that this procedure has some issues in the lower quantiles of the normal distribution. This is ascribed to the MLE process convergence to a boundary point of the parameters space for some bootstrap samples.

Finally, we tried another resampling technique to check if we would achieve consistent results in the POT-GPD and POT-Weibull parameters estimation, namely the Jack-knife resampling. In this technique, if the original dataset of n_u exceedances is employed, the $i - th$ jack-knife sample is defined as:

$$x_{(i)} = (x_1; \ldots; x_{i-1}; x_{i+1}; \ldots; x_{nu}) \quad i = 1; \ldots; n_u. \quad (14)$$

A calculation of this method has been performed with a commercial tool for matrix data manipulation. To compare the prediction accuracy for both resampling technique we plotted the standard deviation of the return level predictions as a function of the return level calculated with (12). As we can see in Fig. 13, the jack-knife resampling provides the smallest

standard deviation for estimates (there is a difference up to 40 times at die level prediction) performed for read disturb post-retention. This result is attributed to a small variation of the new generated samples (the replica datasets differ for a single value). To this extent, Jack-knife resampling technique is not well suited to be used together with the POT approach, since generated replicas are not so different, hence, estimations based on these samples differ slightly and could lead to optimistic predictions.

VII. Conclusion

In this work, we validated the POT methodology as a technique for EVA to be applied on a study case like the read disturb reliability modeling in 3D NAND Flash memories. The effectiveness of the model proven its applicability in die level reliability predictions of the number of errors per CW in an important scenario like the post-retention use case. The methodology could be beneficial for storage system level designers dealing with error mitigation schemes. In future, we plan to apply the methodology to consider other 3D NAND Flash reliability threats and to model extreme events in SSD platforms studied at architectural level.

References

[1] R. Micheloni, S. Aritome, and L. Crippa, "Array architectures for 3-D NAND flash memories," *Proc. IEEE*, vol. 105, no. 9, pp. 1634–1649, Sep. 2017, doi: 10.1109/JPROC.2017.2697000.

[2] A. S. Spinelli, C. M. Compagnoni, and A. L. Lacaita, "Reliability of NAND flash memories: Planar cells and emerging issues in 3D devices," *Computers*, vol. 6, no. 2, p. 16, 2017, doi: 10.3390/computers6020016.

[3] T. A. Marquart, "Solid-state-drive qualification and reliability strategy," in *Proc. IEEE Int. Integr. Rel. Workshop (IIRW)*, South Lake Tahoe, CA, USA, Oct. 2015, pp. 3–6, doi: 10.1109/IIRW.2015.7437056.

[4] N. Mielke *et al.*, "Bit error rate in NAND flash memories," in *Proc. IEEE Int. Rel. Phys. Symp.*, Phoenix, AZ, USA, Apr. 2008, pp. 9–19, doi: 10.1109/RELPHY.2008.4558857.

[5] T. Parnell, N. Papandreou, T. Mittelholzer, and H. Pozidis, "Modelling of the threshold voltage distributions of sub-20nm NAND flash memory," in *Proc. IEEE Global Commun. Conf.*, Austin, TX, USA, Dec. 2014, pp. 2351–2356, doi: 10.1109/GLOCOM.2014.7037159.

[6] N.-J. Wang *et al.*, "Statistical analysis of bit-errors distribution for reliability of 3-D NAND flash memories," in *Proc. IEEE Int. Rel. Phys. Symp. (IRPS)*, Dallas, TX, USA, Apr. 2020, pp. 1–5, doi: 10.1109/IRPS45951.2020.9128993.

[7] C. Zambelli, R. Micheloni, and P. Olivo, "Reliability challenges in 3D NAND flash memories," in *Proc. IEEE 11th Int. Memory Workshop (IMW)*, Monterey, CA, USA, May 2019, pp. 1–4, doi: 10.1109/IMW.2019.8739741.

[8] L. Zuolo, C. Zambelli, R. Micheloni, and P. Olivo, "Solid-state drives: Memory driven design methodologies for optimal performance," *Proc. IEEE*, vol. 105, no. 9, pp. 1589–1608, Sep. 2017, doi: 10.1109/JPROC.2017.2733621.

[9] T. Zhang, *Using LDPC Codes in SSD—Challenges and Solutions*, Flash Memory Summit, Santa Clara, CA, USA, Aug. 2012.

[10] E. F. Haratsch, *LDPC Code Concepts and Performance on High-Density Flash Memory*, Flash Memory Summit, Santa Clara, CA, USA, Aug. 2014.

[11] N. R. Mielke, R. E. Frickey, I. Kalastirsky, M. Quan, D. Ustinov, and V. J. Vasudevan, "Reliability of solid-state drives based on NAND flash memory," *Proc. IEEE*, vol. 105, no. 9, pp. 1725–1750, Sep. 2017, doi: 10.1109/JPROC.2017.2725738.

[12] S. Im and D. Shin, "Flash-aware RAID techniques for dependable and high-performance flash memory SSD," *IEEE Trans. Comput.*, vol. 60, no. 1, pp. 80–92, Jan. 2011, doi: 10.1109/TC.2010.197.

[13] J. Kim, E. Lee, J. Choi, D. Lee, and S. H. Noh, "Chip-level RAID with flexible stripe size and parity placement for enhanced SSD reliability," *IEEE Trans. Comput.*, vol. 65, no. 4, pp. 1116–1130, Apr. 2016, doi: 10.1109/TC.2014.2375179.

[14] Y. Li, P. P. C. Lee, and J. C. S. Lui, "Analysis of reliability dynamics of SSD RAID," *IEEE Trans. Comput.*, vol. 65, no. 4, pp. 1131–1144, Apr. 2016, doi: 10.1109/TC.2014.2349505.

[15] C. Zambelli, A. Marelli, R. Micheloni, and P. Olivo, "Modeling the endurance reliability of intradisk RAID solutions for mid-1X TLC NAND flash solid-state drives," *IEEE Trans. Device Mater. Rel.*, vol. 17, no. 4, pp. 713–721, Dec. 2017, doi: 10.1109/TDMR.2017.2749639.

[16] A. Grossi, L. Zuolo, F. Restuccia, C. Zambelli, and P. Olivo, "Quality-of-service implications of enhanced program algorithms for charge-trapping NAND in future solid-state drives," *IEEE Trans. Device Mater. Rel.*, vol. 15, no. 3, pp. 363–369, Sep. 2015, doi: 10.1109/TDMR.2015.2448108.

[17] C. Zambelli, L. Crippa, R. Micheloni, and P. Olivo, "Points-over-threshold statistics for post-retention read disturb reliability in 3D NAND flash," in *Proc. IEEE Int. Integr. Rel. Workshop (IIRW)*, South Lake Tahoe, CA, USA, 2020, pp. 1–5.

[18] M. R. Leadbetter, "On a basis for 'peaks over threshold' modeling," *Stat. Probab. Lett.*, vol. 12, no. 4, pp. 357–362, 1991, doi: 10.1016/0167-7152(91)90107-3.

[19] K. Ha, J. Jeong, and J. Kim, "A read-disturb management technique for high-density NAND flash memory," in *Proc. 4th Asia–Pac. Workshop Syst.*, 2013, pp. 1–6, doi: 10.1145/2500727.2500743.

[20] C. Zambelli *et al.*, "Characterization of TLC 3D-NAND flash endurance through machine learning for LDPC code rate optimization," in *Proc. IEEE Int. Memory Workshop (IMW)*, Monterey, CA, USA, 2017, pp. 1–4, doi: 10.1109/IMW.2017.7939074.

[21] *Electrically Erasable Programmable Rom (EEPROM) Program/Erase Endurance and Data Retention Test* document JESD22-A117, JEDEC, Arlington, VA, USA, Oct. 2018.

[22] C. Zambelli, P. Olivo, L. Crippa, A. Marelli, and R. Micheloni, "Uniform and concentrated read disturb effects in mid-1X TLC NAND flash memories for enterprise solid state drives," in *Proc. IEEE Int. Rel. Phys. Symp. (IRPS)*, Monterey, CA, USA, 2017, pp. 1–4, doi: 10.1109/IRPS.2017.7936387.

[23] C. Zambelli, R. Micheloni, S. Scommegna, and P. Olivo, "First evidence of temporary read errors in TLC 3D-NAND flash memories exiting from an idle state," *IEEE J. Electron Devices Soc.*, vol. 8, pp. 99–104, 2020, doi: 10.1109/JEDS.2020.2965648.

[24] (2019). *Microsemi PM8609 NVMe2032 Flashtec NVMe Controller*. [Online]. Available: https://www.microsemi.com/product-directory/storage-ics/3687-flashtec-nvme-controllers

[25] N. Papandreou *et al.*, "Characterization and analysis of bit errors in 3D TLC NAND flash memory," in *Proc. IEEE Int. Rel. Phys. Symp. (IRPS)*, Monterey, CA, USA, 2019, pp. 1–6, doi: 10.1109/IRPS.2019.8720454.

[26] F. Wang *et al.*, "Lateral charge migration induced abnormal read disturb in 3D charge-trapping NAND flash memory," *Appl. Phys. Exp.*, vol. 13, no. 5, Apr. 2020, Art. no. 054002, doi: 10.35848/1882-0786/ab8729.

[27] Y. Cai, Y. Luo, S. Ghose, and O. Mutlu, "Read disturb errors in MLC NAND flash memory: Characterization, mitigation, and recovery," in *Proc. IEEE/IFIP Int. Conf. Depend. Syst. Netw.*, Rio de Janeiro, Brazil, Jun. 2015, pp. 438–449, doi: 10.1109/DSN.2015.49.

[28] V. Taranalli, H. Uchikawa, and P. H. Siegel, "On the capacity of the beta-binomial channel model for multi-level cell flash memories," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 9, pp. 2312–2324, Sep. 2016, doi: 10.1109/JSAC.2016.2603660.

[29] Y. Luo, S. Ghose, Y. Cai, E. F. Haratsch, and O. Mutlu, "Improving 3D NAND flash memory lifetime by tolerating early retention loss and process variation," 2018. [Online]. Available: arXiv:1807.05140.

[30] Y. Cai, E. F. Haratsch, O. Mutlu, and K. Mai, "Threshold voltage distribution in MLC NAND flash memory: Characterization, analysis, and modeling," in *Proc. Design Autom. Test Eur. Conf. Exhibit. (DATE)*, Grenoble, France, 2013, pp. 1285–1290, doi: 10.7873/DATE.2013.266.

[31] C. Stander, "Analysis of extreme events in the coastal engineering environment," M.S. thesis, Dept. Appl. Math. Stellenbosch Univ., Stellenbosch, South Africa, Dec. 2015.

[32] N. Teena, V. S. Kumar, K. Sudheesh, and R. Sajeev, "Statistical analysis on extreme wave height," *Nat. Hazards J. Int. Soc. Prevent. Mitigation Nat. Hazards*, vol. 64, no. 1, pp. 223–236, Oct. 2012, doi: 10.1007/s11069-012-0229-y.

[33] S. Coles, *An Introduction to Statistical Modeling of Extreme Values*. London, U.K.: Springer, 2001.

[34] E. Merran, N. Hastings, and B. Peacock, *Statistical Distributions*, 2nd ed. New York, NY, USA: Wiley, 1993.

IEEE Proof